Contemporary Educational Psychology 54 (2018) 184–198

Contents lists available at ScienceDirect

Contemporary Educational Psychology

journal homepage: www.elsevier.com/locate/cedpsych

Empirical study

Scaffolding scientific thinking: Students' evaluations and judgments during Earth science knowledge construction

Doug Lombardi[a,*], Janelle M. Bailey[a], Elliot S. Bickel[b], Shondricka Burrell[a]

[a] Department of Teaching & Learning, Temple University, United States
[b] George Washington Carver Engineering & Science High School, United States

ABSTRACT

Critical evaluation underpins the practices of science. In a three-year classroom-based research project, we developed and tested instructional scaffolds for Earth science content in which students evaluate lines of evidence with respect to alternative explanations of scientific phenomena (climate change, fracking and earthquakes, wetlands and land use, and formation of Earth's Moon). The present paper documents a quasi-experimental study where high school Earth science students completed these instructional scaffolds, including an explanation task scored for evaluative levels (erroneous, descriptive, relational, and critical), along with measures of plausibility reappraisal and knowledge. Repeated measures analyses of variance reveal significant increases in plausibility and knowledge scores for students completing instructional scaffolds that promoted students' evaluations about the connections between lines of evidence and two alternative explanations, whereas evaluations about connections between lines of evidence and only one alternative show no change in scores. A structural equation model suggests that students' evaluation may influence post instructional plausibility and knowledge. The results of this study demonstrate that students' active evaluation of scientific alternatives and explicit reappraisal of plausibility judgments can support deeper learning of Earth science content.

1. Introduction

Scientific literacy involves both knowing *what* scientists know and knowing *how* scientists know what they know. Recent science education reform efforts capture these two essential components (i.e., the *what* and the *how* of scientific knowledge) in a three-dimensional learning framework that intertwines scientific practices, crosscutting concepts, and disciplinary core ideas (National Research Council [NRC], 2012; NGSS Lead States, 2013). In this framework, evaluative processes act as a central hub linking the scientific activities of empirical inquiry and constructing explanations. Although the framework embeds reasoning throughout, evaluation as argument, critique, and analysis is central to scientific thinking and knowledge construction (NRC, 2012).

Evaluation often follows a dynamic and iterative cycle in the scientific enterprise. For example, some climatologists construct explanatory and predictive models representing Earth's atmosphere, and then collect empirical data to calibrate these models. Evaluation of connections between lines of evidence (e.g., sea surface temperatures) and scientific explanations (e.g., the interdependence of oceans and atmosphere) could lead to subsequent model refinements and validation with additional empirical data. Conant (1951) describes this dynamic and evaluative process as the speculative enterprise of science, where scientific knowledge construction is complex and requires mature, evaluative thinking (Kuhn & Pearsall, 2000). But such thinking may be difficult for students to learn and for teachers to teach (see, for example, Erduran & Dagher, 2014; Klopfer, 1969). Because of this difficulty, instructional scaffolds may be required to help students learn how to critically evaluate connections between evidence and explanations (Greene, Hutchison, Costa, & Crompton, 2012; Li et al., 2016) and construct scientifically accurate knowledge (Duschl, 2008; Sandoval & Reiser, 2004).

Our recent classroom-based research project focused on developing and testing instructional scaffolds—called Model-Evidence Link (MEL) diagrams[1]—that facilitate students' evaluations and judgments during knowledge construction (Fig. 1). Our project concentrated on the Earth science domain, which includes many topics that are challenging for

* Corresponding author at: 450 Ritter Hall, 1301 Cecil B. Moore Avenue, Philadelphia, PA 19122, United States.
E-mail address: doug.lombardi@temple.edu (D. Lombardi).

[1] A team of researchers at Rutgers University developed the original structure and mode of the MEL (see Chinn & Buckland, 2012, for an overview). Lombardi, Sinatra, and Nussbaum (2013) and Lombardi, Bickel, Bailey, and Burrell (2018) adapted and expanded upon this design to fit within their theoretical framework positing the dynamic relations between evaluation, plausibility, and knowledge. The MEL is an instructional scaffold that facilitates students' evaluations about the connections between multiple lines of evidence and alternative explanations about a phenomenon (e.g., causes of current climate change).

https://doi.org/10.1016/j.cedpsych.2018.06.008

Available online 23 June 2018
0361-476X/

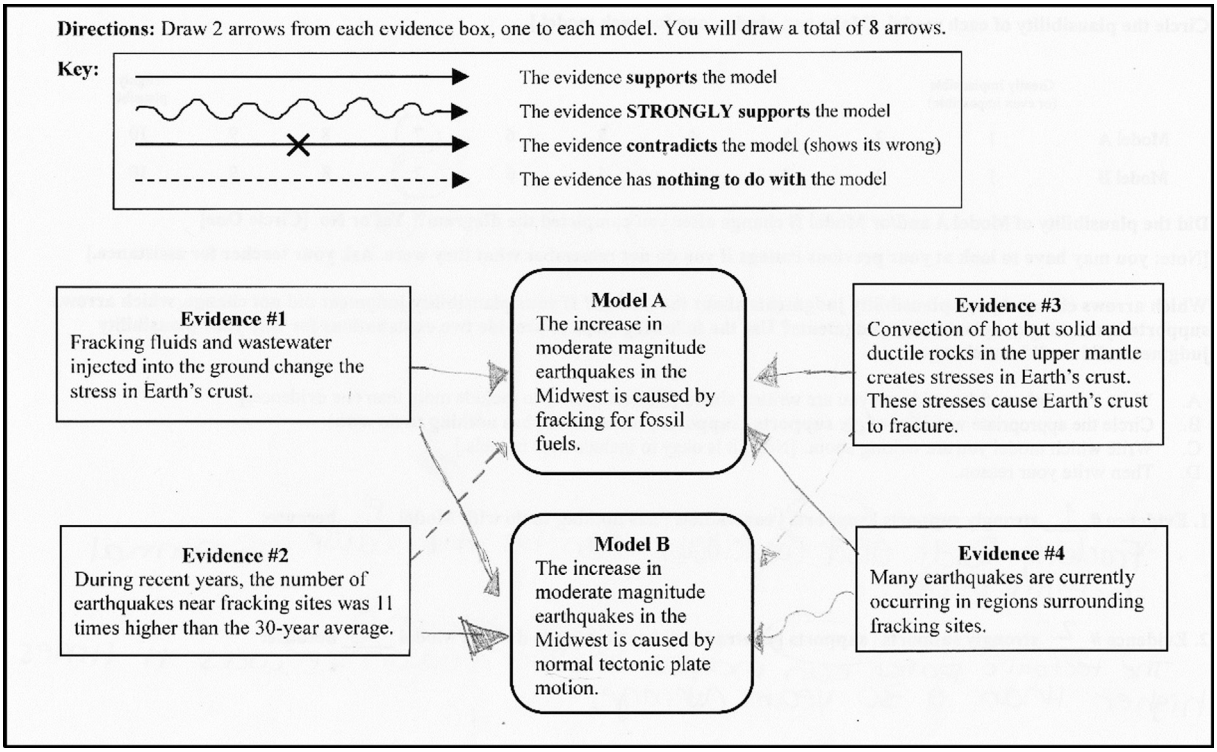


Fig. 1. A student example of the fracking Model-Evidence Link (MEL) diagram.

students because: (a) the underlying scientific principles are complex, (b) the processes frequently occur over very long time and large spatial scales, and (c) students have difficulty understanding how scientifically accurate explanations are constructed. Furthermore, some topics in Earth science are particularly salient because they concern issues of great local, regional, and global importance (e.g., climate change; see, for example, Sadler, Klosterman, & Topcu, 2011). Therefore, investigating scaffolds (e.g., the MEL) that help students think more scientifically about Earth science—specifically within a classroom context—may be both relevant and useful to systematically understand contemporary learning environments (Barab & Squire, 2004). Our project, and specifically the present study, comparatively examines MELs in authentic secondary classroom settings with the goal of gauging how Earth science students can deepen their knowledge about natural phenomena through scientific evaluations and judgments.

The present study examines the MEL in comparison to two other scaffolds. Lombardi, Nussbaum, and Sinatra (2016) argued that evaluative comparisons of alternative explanations could facilitate students' knowledge construction through increased cognitive engagement. Therefore, we specifically compared the MEL, where students evaluate connections between lines of evidence and two alternative explanations (i.e., the scientific alternative vs. another alternative), to the Mono-MEL, where students evaluate connections between lines of evidence and only one explanation (i.e., the scientific alternative). We also compared the MEL, where students evaluate connections diagrammatically, to the Model-Evidence Link Table (MET), where students evaluate connections using tables and letter codes. In subsequent sections, we elaborate further on our justifications for comparing these three scaffolds.

We built our project on a theoretical perspective that posits the following: learners may construct scientifically accurate knowledge through a process of generating explicit evaluations about scientific evidence and reappraising their plausibility judgments about explanations (Lombardi, Nussbaum et al., 2016). This perspective has both philosophical foundations (Rescher, 2009; Salmon, 1994), and empirical bases in educational, developmental, and cognitive psychology (Chinn & Brewer, 2001; Collins & Michalski, 1989; Connell & Keane, 2006; Dole & Sinatra, 1998; Kuhn & Pearsall, 2000; Nussbaum, 2011) and science education research (Braaten & Windschitl, 2011; Chi, 2005; Chinn & Brewer, 1993). Our discussion below highlights the extant literature supporting this theoretical perspective, as well as recent empirical work examining ways to promote scientific thinking and knowledge construction through more critical evaluations and judgments, and situates this perspective within the context of the present study.

1.1. Evaluation, plausibility, and knowledge

All scientific practices emerge from "processes of perpetual evaluation and critique that support progress in explaining nature" (Ford, 2015, p. 1043), and recent science education reform efforts call for students to engage in the scientific practices to help them achieve college- and career-readiness (NRC, 2012). To effectively participate in these practices, students should cognitively evaluate scientific evidence and "plausible explanation[s] for an observed phenomenon that can predict what will happen in a given situation" (NRC, 2012, p. 67). The scientific community also compares the plausibility of alternative explanations when constructing scientific models and theories. Yet within the context of certain Earth science phenomena (e.g., climate change and hydraulic fracturing, aka "fracking"), scientists may generate explanations that seem implausible to students. In contrast, alternative lay explanations about such phenomena – such as the notion that increasing amounts of energy received from the Sun are the cause of current climate change – may seem more plausible than scientific ones. Students may consider this lay explanation more plausible than the scientific explanation that human activities are the cause of current climate change. This difference in judgments about what explains a phenomenon is what Lombardi, Sinatra, and Nussbaum (2013) call a "plausibility gap."

Plausibility judgments may be associated with critical and scientific thinking. For example, Beyer (1995) says that questioning the plausibility of explanations is one characteristic of skepticism, a disposition of

critical thinkers. Differentiating between evidence that supports the truthfulness of a claim, and theory that supports the plausibility of a claim, is also a characteristic of those who are developing scientific thinking skills (Kuhn, 1999). By examining a theory's potential truthfulness, plausibility judgments used in a critical mode may be evaluative. Such critical evaluations about the plausibility of explanations are also fundamentally linked to an individual's knowledge (Willingham, 2008), based on the presupposition that plausibility judgments are tentative in nature and may contribute to knowledge construction (Lombardi, Nussbaum et al., 2016). Although explicit and critical evaluations of novel explanations may influence appraisals of plausibility, people's implicit perceptions and biases may activate cognitive processes that are not reflective and purposeful. Lombardi, Nussbaum et al. (2016) speculate that plausibility judgments often form implicitly and without much thought, thereby necessitating that students be more critically evaluative when judging the plausibility of scientific explanations.

Plausibility judgments have also long been theoretically implicated as one of many important factors in the process of science learning (see, for example, Chinn & Brewer, 1993; Dole & Sinatra, 1998; Kapon & diSessa, 2012; Posner, Strike, Hewson, & Gertzog, 1982), but until recently, almost no empirical research has validated the importance of plausibility in knowledge construction and reconstruction (see Lombardi, Nussbaum et al., 2016, for a detailed philosophical, empirical, and theoretical review). Although recent research (see, for example, Lombardi, Bickel, Bailey, & Burrell, 2018) shows that importance of plausibility judgments in knowledge construction, Lombardi, Nussbaum et al. (2016) state that students are only most "likely to engage with ideas that are perceived to have [both] high plausibility and cognitive utility" (p. 49). Indeed, other factors, such as commitment-based social group membership, could override increased plausibility (Dole & Sinatra, 1998), which makes the process of constructing conceptions consistent with scientific understanding difficult (Chi, 2005). Lombardi, Nussbaum et al. (2016) recently proposed a theoretical model that posits initial plausibility judgments might be reappraised through the process of being critically evaluative (i.e., plausibility reappraisal may elevate initial plausibility judgments from regimes of low/implicit evaluation to high/explicit evaluation). Reappraisal, in turn, may be a component of constructing scientifically accurate knowledge, but only if the plausibility judgment is now considered greater than the plausibility of preexisting and/or alternative conceptions, and only if other factors, such as personal stake in the outcome, do not strongly override the plausibility judgment.

1.2. Scientific thinking through evaluation

Students may be naturally curious about scientific topics, but they are not necessarily evaluative as they consider hypotheses and theories constructed by scientists. The process of evaluation can involve judgments about the relationship between evidence and alternative explanations of a particular phenomenon (McNeill, Lizotte, Krajcik, & Marx, 2006). Therefore, when students are more critical in their evaluation of scientific knowledge they will seek to weigh the strengths and weaknesses in the connection between evidence and explanations, and gauge how well evidence potentially supports both an explanation (e.g., an argument, a scientific model) and its plausible alternatives (e.g., a counterargument, a contrary hypothesis). Students' evaluations should also reflect on the process of knowledge construction to promote deeper understanding (Mason, Ariasi, & Boldrin, 2011). When students model practices used by scientific experts they may cognitively reflect and evaluate in a manner similar to scientists (Duschl, Schweingruber, & Shouse, 2007). Students who engage in reflective evaluation understand that scientific knowledge emerges from collaborations that are constructive, critical, and open to revision (Nussbaum, 2008). Because students may not be critically reflective when engaging in collaborative knowledge construction, they may need instructional scaffolds to evaluate the quality of explanations (Gijlers & de Jong, 2009; Kyza, 2009; Metz, 2004; Nussbaum & Edwards, 2011). Instructional scaffolds that facilitate students' coordination of lines evidence with alternative explanations (e.g., the MEL) hold some promise for deepening students' science learning (Chinn & Buckland, 2012; Lombardi, Danielson, & Young, 2016). Such scaffolds may be particularly useful because the process of weighing the connections between lines of evidence and more than one explanation may help students become more scientific and critical in their evaluations, which in turn could promote plausibility reappraisal and deeper knowledge construction (Lombardi et al., 2018, 2013, Lombardi, Danielson et al., 2016).

1.3. Relating evaluation and plausibility

Researchers have implicated plausibility judgments in facilitating co-construction of knowledge in discourse associated with collaborative argumentation (Duschl et al., 2007; Nussbaum, 2011). Researchers have also proposed that plausibility may be an important judgment involved in construction of scientifically accurate knowledge (Dole & Sinatra, 1998; Pintrich, Marx, & Boyle, 1993; Posner et al., 1982). Lombardi, Nussbaum et al.'s (2016) theoretical model describes how plausibility judgments often may be formed through automatic cognitive processes (i.e., cognitive activities that require very little attentional capacity; LaBerge & Samuels, 1974; Stanovich, 1990). However, explicit prompting (i.e., via instruction) may facilitate reappraisal of these implicit plausibility judgments toward a more scientific stance. Such instruction may be particularly relevant for complex and abstract scientific topics (e.g., climate change), where a gap exists between what students and scientists find plausible.

Empirical research has revealed that a plausibility gap exists for the topic of global climate change among middle school students (Lombardi et al., 2013), undergraduate students (Lombardi & Sinatra, 2012; Lombardi, Danielson et al., 2016), and elementary and secondary science teachers (Lombardi & Sinatra, 2013). To address this gap, Lombardi et al. (2013) developed a MEL for the topic of climate change. Grade 7 students who used this MEL experienced significant shifts in both plausibility and knowledge toward the scientifically accepted model of human-induced climate change. These students also retained their knowledge gains six months after instruction. In comparison, grade 7 students at the same school and taught by the same teachers did not experience plausibility or knowledge shifts when experiencing another instructional activity designed to promote scientific inquiry and deeper understanding of climate change (Smith, Southard, & Mably, 2002). This comparison activity asked students to construct their own explanations based on evidence, rather than weigh evidence between two competing models of climate change (i.e., as the treatment activity did). Lombardi et al. (2013) speculated that the students' plausibility reappraisal—a skill that is important for understanding the development of scientific knowledge (Duschl et al., 2007; Hogan & Maglienti, 2001)—was related to the MEL's ability to facilitate students' critical evaluation. Plausibility reappraisal, in turn, may have promoted the students' enduring knowledge gains (Erduran & Dagher, 2014).

1.4. Linking scientific practices to evaluation and plausibility reappraisal

Recent empirical research closely examining student work on the MEL activities shows that students engage in various levels of evaluation when considering alternative explanations about Earth and space science phenomena and that these evaluation levels are significantly related to plausibility appraisals and knowledge about the phenomena (Lombardi, Brandt, Bickel, & Burg, 2016, Lombardi et al., 2018). Specifically, high school students shifted plausibility toward scientifically accepted explanations and increased their knowledge about relevant Earth science topics after participating in MEL activities. Greater levels of evaluation were related to plausibility shifts and knowledge increases, as shown by structural equation modeling. Effect sizes were

small to large, depending upon topic and instructional context. For example, on one hand, there was a large effect size where combined knowledge scores increased over time. On the other hand, individual classroom settings revealed large effect sizes for some topics (e.g., the connections between fracking and earthquakes) and small effect sizes for others (e.g., the importance of wetland resources). These findings support the idea that MEL activities moved students to cognitively engage in practices that helped them think more scientifically. Specifically, students learned "that alternative interpretations of scientific evidence can occur, that such interpretations must be carefully scrutinized, that the plausibility of the supporting evidence must be considered, [and] …that predictions or explanations can be revised on the basis of seeing new evidence or of developing a new model that accounts for the existing evidence better than previous models did" (NRC, 2012, pp. 251–252). However, we still wondered if the MEL was particularly effective at promoting evaluation and plausibility reappraisal, or if students would perform as well, or better, when engaging in other, similar tasks. This general question motivated the present study.

1.5. The present study

The present study represents the culmination of a three-year classroom-based research project. This project's overall purpose was to design and test instructional scaffolds, based on Lombardi, Nussbaum et al.'s (2016) theoretical perspective, that (a) facilitate students' scientific evaluations about the connections between lines of evidence and alternative explanations, (b) promote shifts in their plausibility judgments about scientific explanations, and (c) deepen students' scientifically accurate knowledge about Earth science phenomena. The project was a collaboration between master teachers and researchers per Anderson and Shattuck's (2012) guidance. Specifically, we partnered with teachers in identifying the initial problem, designing and constructing the interventions, and creating publications for fellow practitioners. In the project's first year, the team designed three MEL diagrams and associated materials, covering the topics of (a) fracking and earthquakes, (b) wetlands and land use, and (c) formation of Earth's Moon.

We added these three to the existing climate change MEL developed by Lombardi et al. (2013). We chose these four topics (causes of current climate change, relations between fracking and earthquakes, use of wetlands, and formation of Earth's Moon) because each has multiple plausible explanatory models (a scientifically accepted and an alternative) that students could evaluate. Furthermore, these four represent a wide range of topics that might be covered in a typical high school Earth science scope and sequence. In authentic classroom settings, our team conducted pilot testing of these four instructional scaffolds during the initial year. We revised the scaffolds at the end of the first year based on feedback from the teachers. During the second year of the project, we again tested the full suite of MELs in four school settings (see Lombardi et al., 2018 for details on the second-year testing and study) and made final revisions to all materials in preparation for the project's third year.

The present study is associated with a quasi-experimental design phase of the project (third year). In this phase, we examined the finalized MEL materials with two other comparison activities: the MET, which used a table format in place of a diagram, and the Mono-MEL diagram, which used a diagram format similar to the MEL but with only one model, the scientific explanation. We conducted the present study in authentic school settings that had not been previously involved in the project, and specifically examined two research questions:

1. How do instructional scaffolds promoting evaluation of alternatives (i.e., the MEL and MET) compare to one that does not (i.e., the Mono-MEL), specifically in the shifting plausibility judgments and changing knowledge toward scientifically accurate understanding?
2. What are the relations between evaluation, plausibility, and knowledge, and how do instructional scaffolds promoting evaluation of the connections between lines of evidence and alternative explanations facilitate plausibility reappraisal and knowledge construction over the course of a school year?

The novel aspect of the present study was our use of comparison activities that provided a robust test of the potential effectiveness of scaffolds designed to promote more critical evaluations (e.g., the MEL). Whereas one previous study has measured the effectiveness of the MEL against more traditional science instruction (i.e., generating evidence-based explanations; Lombardi et al., 2013), the current study compared the MEL, where students evaluated connections between lines of evidence and two alternative models, to the Mono-MEL, where students evaluated lines of evidence to only one model (i.e., the scientific explanation). Lombardi et al. (2013) suggested that the appreciable instructional advantage of the MEL (i.e., over the more traditional activity) was due to the MEL's structure, where students evaluated connections between lines of evidence and two alternative explanations. In short, evaluations involving alternative explanations may be deeper and more critical than evaluations considering only one explanation. Therefore, we hypothesized that the MEL would be more effective than the Mono-MEL in both facilitating plausibility reappraisal and deepening knowledge. Because we were curious about the graphical nature of the MEL diagram, the present study also compared the MEL to the MET, which also asked students to evaluate connections between lines of evidence and two alternative models but did so in a tabular format. We hypothesized that MEL may have slightly more favorable outcomes (i.e., in terms of plausibility reappraisal and knowledge changes) because the directionality of the arrows might reinforce the causal relationship between lines of evidence and the explanatory models. In the MET, there is no visual cue that the evidence may be causally linked to the model (i.e., the code is only a pairing), and the lack of such relationship could influence evaluations and plausibility judgments (Chinn & Brewer, 2001).

2. Methods

2.1. Setting and participants

The present study involved high school (grades 9–12) students from two schools, one located in a large urban district in the Southwest US, and one located in medium-sized suburban district in the Mid-Atlantic US. Both schools were located in states that have adopted the Next Generation Science Standards (NGSS). Students involved in the present study were enrolled in Earth science classes taught by Ms. Rodgers (Southwest US school) and Ms. Williams (Mid-Atlantic US school; both teacher names are pseudonyms). Because of the nature of the research questions (i.e., looking at how instruction over the course of a school year affects evaluations, plausibility judgments, and knowledge construction), we only included students who completed all instructional tasks and measures. Furthermore, we only included those students who provided assent to participate in the research and whose parents provided parental consent. Sixty-four students met these requirements and were participants in the present study, with just over half being in Ms. Rodgers' classes ($n = 34$) and the remainder being in Ms. Williams' classes ($n = 30$). About 53% ($n = 18$) of Ms. Rodgers' students indicated they were male, 41% ($n = 14$) indicated they were female, and 6% ($n = 2$) did not indicate gender. In Ms. Rodgers' classes, about 41% ($n = 14$) indicated they were Hispanic or Latino, 12% ($n = 4$) indicated they were White, 12% ($n = 4$) indicated they were Black or African American, 17% ($n = 6$) indicated they were Asian, Native Hawaiian, or other Pacific Islander, 12% ($n = 4$) indicated they were two or more races/ethnicities, and 6% ($n = 2$) did not indicate race/ethnicity. About 73% ($n = 22$) of Ms. Williams' students indicated they were male and 27% ($n = 8$) indicated they were female. In Ms. Williams' classes, about 23% ($n = 7$) indicated they were Hispanic or Latino, 57% ($n = 17$)

Table 1
Summary of models and evidence statements for each Model-Evidence Link (MEL) activity.

Topic	Model		Evidence statements
	Scientific	Alternative	
Climate Change	Our current climate change is caused by increasing amounts of gases released by human activities. (Model A)	Our current climate change is caused by increasing amounts of energy released from the Sun. (Model B)	#1: Atmospheric greenhouse gas concentrations have been rising for the past 50 years. Human activities have led to greater releases of greenhouse gases. Temperatures have also been rising during these past 50 years. #2: Solar activity has decreased since 1970. Lower activity means that Earth has received less of the Sun's energy. But, Earth's temperature has continued to rise. #3: Satellites are measuring more of Earth's energy being absorbed by greenhouse gases. #4: Increases and decreases in global temperatures closely matched increases and decreases in solar activity before the industrial revolution.
Fracking	The increase in moderate magnitude earthquakes in the Midwest is caused by fracking for fossil fuels. (Model A)	The increase in moderate magnitude earthquakes in the Midwest is caused by normal tectonic plate motion. (Model B)	#1: Fracking fluids and wastewater injected into the ground change the stress in Earth's crust. #2: During recent years, the number of earthquakes near fracking sites was 11 times higher than the30-year average. #3: Convection of hot but solid and ductile rocks in the upper mantle creates stresses in Earth's crust. These stresses cause Earth's crust to fracture. #4: Many earthquakes are currently occuring in regions surrounding sites.
Wetlands	Wetlands provide ecosystem services that contribute to human welfare and help sustain the biosphere.[a] (Model A)	Wetlands are a nuisance to humans and provide little overall environmental benefit. (Model B)	#1: Wetlands play a role in the global cycles of carbon, nitrogen, and sulfur. Wetlands change these nutrients into different forms necessary to continue their global cycles. #2: Flooding is a natural occurrence in low-lying areas and wetlands are places where floodwaters can collect. #3: Wetlands contribute 70 percent of global atmospheric methane from natural sources. #4: Many wetlands are located in rapidly developing areas of the country.
Moon	The Moon formed after a large object collided with Earth and material from both combined to create the Moon. (Model B)	The Moon was an object that came from elsewhere in the solar system and was captured by Earth's gravity. (Model A)	#1: Earth's average density is higher than the Moon's. The density of Earth's crust is a little less than the Moon's, but Earth's density increases toward the core. #2: Simulations of other star systems show that planets form when smaller objects collide. #3: The Moon's orbit around Earth is tilted compared to the planets' orbits around the Sun. #4: Earth is about 35% iron, most of which is in the core. The Moon has very little iron.

[a] Although a socio-scientific topic, the wetlands MEL asks students to make judgments about "value" models rather than scientific explanatory models. The "scientific" model in this case is that with which most environmental scientists agree.

indicated they were White, 10% (n = 3) indicated they were Black or African American, 3% (n = 1) indicated they were two or more races/ethnicities, and 7% (n = 2) did not indicate race/ethnicity.

2.2. Materials

2.2.1. Instructional scaffolds

Our project team has previously detailed development of and research about the four MEL activities (see Lombardi et al., 2018, 2013), including classroom guidance and implementation instructions in alignment with the high school Earth science NGSS performance expectations. However, to provide context for the present study, we highlight some of the important features of the MEL below. We also provide information about the comparison activities used in the present study: the MET and Mono-MEL. In an earlier study, Lombardi et al. (2013) compared the climate change MEL to instructional materials that were consistent with science education reform efforts of the middle and late 1990s and early 2000s (National Research Council, 1996). Lombardi et al. (2013) found the MEL activities significantly outperformed these other materials. However, when our project team considered what would be fair comparisons to the MEL for the present study, we wanted to ensure that the comparison tasks were also aligned with the NGSS and fully supported current curricular goals for students' learning. Therefore, we constructed two comparison activities that represent only slight modifications to the MEL diagram and involved almost all of the same instructional materials.

2.2.1.1. The MEL. Our MEL diagrams consist of four lines of evidence and two alternative explanations about a phenomenon (e.g., increase in moderate magnitude earthquakes in the Midwest US). We used major lines of evidence that scientists have collected about the four topics and constructed simple and declarative one- to three-sentence summary statements highlighting each evidence line (see Table 1 for a complete listing of the lines of evidence for each topic). For example, the fracking lines of evidence included information both about the processes behind fracking and about the occurrence of earthquakes: the effect that fracking injection has on friction in Earth's crust; the changes in frequency of earthquakes near fracking sites; the natural processes that cause earthquakes; and the amount of force exerted on Earth's crust during fracking. We also constructed one-page "evidence texts" for each evidence line. These texts included diagrams and tables to elaborate on the evidence statements, and teachers encouraged participants to use these one-page texts in completing the activity. Our project's advisory board, which included two geoscientists and two educational psychologists, checked each line of evidence and corresponding evidence texts for both face and content validity.

Each MEL diagram also shows two alternative explanatory models about a particular phenomenon (Table 1). One of the models is scientifically accepted and one is an alternative model not accepted by the relevant scientific community (e.g., climatologists). Each MEL presents the alternatives as "Model A" or "Model B," with no indication of the validity of either. To complete the MEL diagram, participants drew arrows of different types between each evidence statement and each alternative model. These types of arrows indicated participants' evaluations about how well a line of evidence supported a model, where a straight line arrow meant that the evidence supported a model, a squiggly line arrow meant the evidence strongly supported a model, a dotted line arrow meant the evidence had nothing to do with the model, and a straight line arrow with an "X" through it meant that the evidence contradicted a model. Participants drew a total of eight arrows to construct each MEL diagram (Fig. 1).

In drawing the arrows, participants' evaluations about the connections between lines of evidence and explanations may have facilitated their reappraisals of the plausibility of each model. Participants probably made their initial plausibility judgments (made prior to drawing the arrows, see procedures below) via implicit processing (i.e., low awareness and low cognitive effort), which is a default cognitive mode (Lombardi, Nussbaum et al., 2016; Stanovich, 2010). Further, these initial and implicit plausibility judgments were probably influenced by a variety of factors that were not reflective of scientific thinking and evaluation (Lombardi, Nussbaum et al., 2016). Therefore, by having students be explicitly evaluative about evidence and model connection criteria (the evidence strongly supports, supports, has nothing to do with, or contradicts the model), the MEL was designed to facilitate reappraisal of participants' initial plausibility judgments about the two explanatory models per Lombardi, Nussbaum et al.'s (2016) theoretical model. Although participants may have used additional criteria to judge the models (e.g., internal and external consistency), the MEL is specifically designed to help participants evaluate how well lines of evidence support two explanatory models.

2.2.1.2. The MET. The MET is quite similar to the MEL in that the activity uses the same lines of evidence, including the same evidence texts, and the same two alternative explanations for each topic. However, instead of drawing different types of arrows on a figure, participants filled in a table with codes (Fig. 2). The research team that created the form and structure of the MEL also recently created the MET, with the idea of using a table format rather than a figure

If you worked with other students, their name(s):_____

Directions: Use the following codes to indicate how well each evidence supports each model. You should put a code into each blank table cell.

Key:

S = The evidence **supports** the model
SS = The evidence **STRONGLY supports** the model
C = The evidence **contradicts** the model (shows its wrong)
N = The evidence has **nothing to do with** the model

	Model A The increase in moderate magnitude earthquakes in the Midwest is caused by fracking for fossil fuels.	**Model B** The increase in moderate magnitude earthquakes in the Midwest is caused by normal tectonic plate motion.
Evidence #1 Fracking fluids and wastewater injected into the ground change the stress in Earth's crust.	C	N
Evidence #2 During recent years, the number of earthquakes near fracking sites was 11 times higher than the 30-year average.	S	N
Evidence #3 Convection of hot but solid and ductile rocks in the upper mantle creates stresses in Earth's crust. These stresses cause Earth's crust to fracture.	N	SS
Evidence #4 Many earthquakes are currently occurring in regions surrounding fracking sites.	S	C

Fig. 2. A student example of the fracking Model-Evidence Link table (MET).

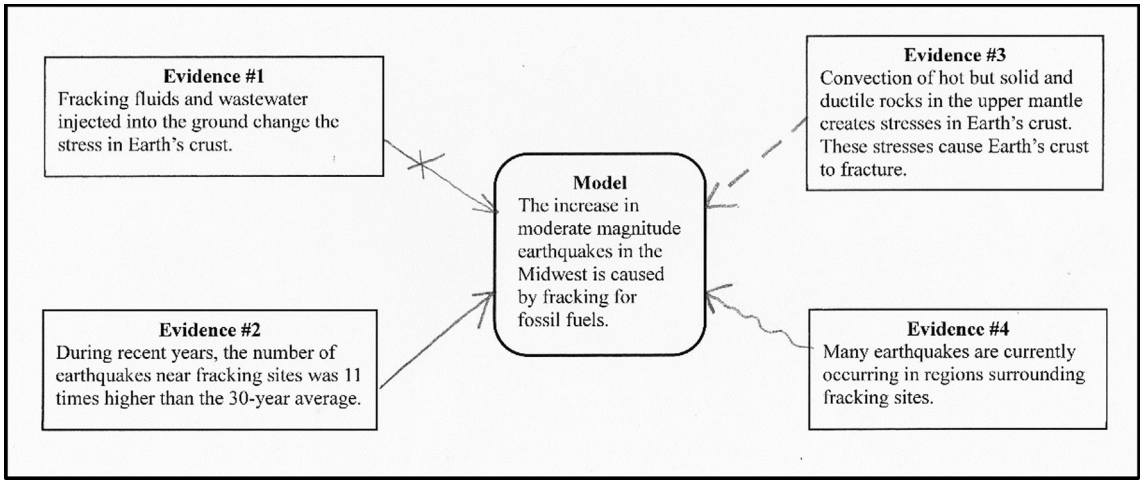


Fig. 3. A student example of the fracking mono Model-Evidence Link (mono-MEL) diagram.

(Rinehart, Golan Duncan, Chinn, Atkins, & DiBenedetti, 2016). For the present study, we slightly modified their MET version to include codes rather than arrows (i.e., evidence strongly supports a model = SS; evidence supports a model = S; evidence contradicts a model = C; and evidence has nothing to do with a model = N). Students completed the MET in the same fashion as the MEL (i.e., by simultaneous comparison of connections between lines of evidence and alternative explanations); however, the MET had a tabular format and the MEL had a graphical format.

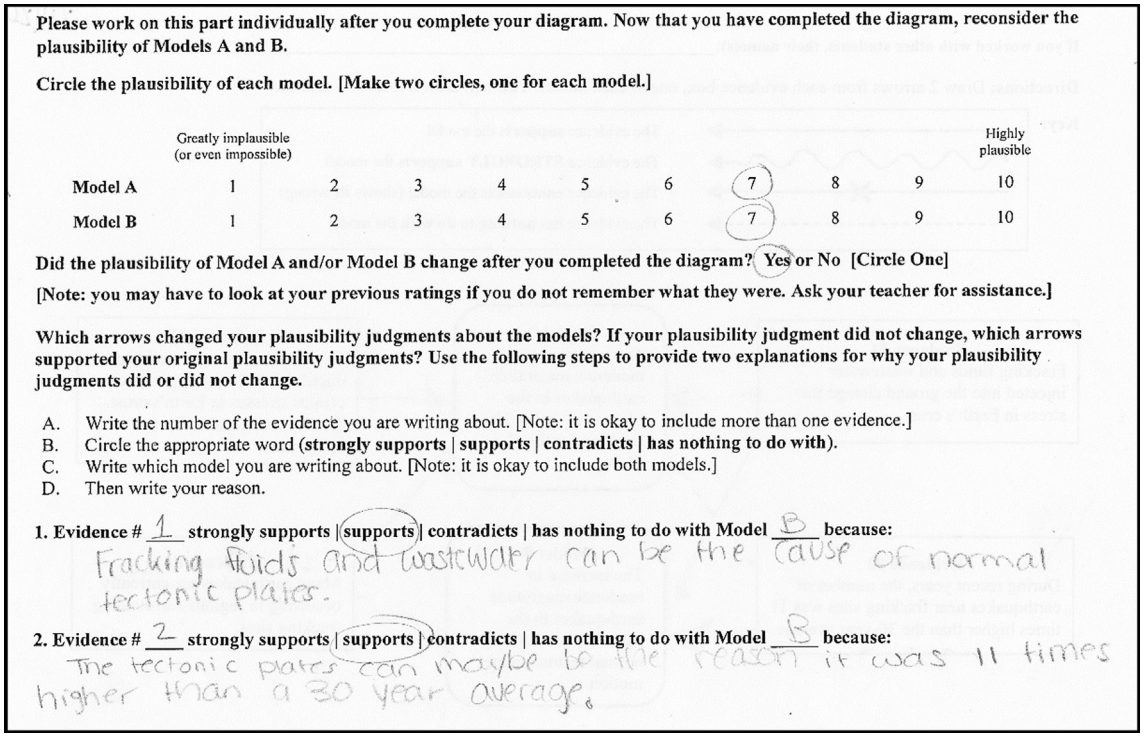
2.2.1.3. The Mono-MEL. The Mono-MEL incorporates the same figure design as the MEL, but only presents one explanation about a phenomenon (i.e., the scientifically accepted explanatory model; Fig. 3). As with the MEL and MET activities, all four lines of evidence, and associated evidence texts, were identical in the Mono-MEL activities. We hypothesized the Mono-MELs would not have as favorable outcomes as the MEL and MET because participants would be considering only one alternative. In an earlier study, Lombardi et al. (2013) speculated that the reason the climate change MEL performed appreciably better than another comparison activity was that students were simultaneously evaluating connections between lines of evidence and alternative explanations with the MEL, but only considering one explanation with the comparison activity.

2.2.2. Explanation task and students' evaluations

After completing either the MEL, MET or Mono-MEL, the "explanation task" (Fig. 4) prompted participants to write about either two or three of the links that they drew on the diagram (MEL and Mono-MEL) or coded in the table (MET). We scored these tasks based on students' levels of evaluation expressed in their written responses. Note that we scored only the explanations, not the diagrams or tables themselves. The explanation task specifically asked participants to describe links between lines of evidence and model(s) that they considered important or interesting. Using a sentence prompt for each explanation, participants wrote down the model (MEL and MET only) and evidence number that they chose to discuss, as well as the evidence-to-model connection strength (i.e., strongly supports, supports, contradicts, or has nothing to do with) they drew on the diagram or wrote in the table. This preface served as the beginning of participants' written explanations, next prompting evaluation with the word "because." For example, one participant's written explanation from the fracking MEL read, "[Evidence 4 strongly supports Model A because…] Most earthquakes occurs [*sic*] near a fracking site which may tell us that fracking causes earthquakes" (note, the section in brackets is part of a sentence frame given to students, with underlined portions filled in by this student). In scoring the explanation tasks, we used a rubric that emerged from a qualitative content analysis by Lombardi, Brandt et al. (2016) in a previous study involving the climate change MEL. Lombardi, Brandt et al. (2016) developed four categories of explanations that drew on the frameworks of both Driver, Leach, Millar, and Scott (1996) and Dole and Sinatra (1998). The categories established four well-defined levels of evaluation to represent the accuracy and elaboration present in participants' responses.

These four different types of evaluations reflect both (a) analyses about the strength of the connections between lines of evidence and explanation and (b) related conceptual understandings. Although Lombardi, Brandt et al. (2016) discuss a detailed qualitative analysis that supports these four levels, the following highlights student responses that relate to each level to give the reader more context. The first category of participants' evaluations, called *erroneous evaluations*, described written responses that represented an incorrect determination about a model-evidence link. Participants who made an erroneous evaluation demonstrated an inability to make a legitimate connection between a line of evidence and model, perhaps from a lack of attention or understanding. Erroneous evaluations prevent deeper comprehension and evaluation from occurring. For example, one student claimed that fracking evidence #1 supports Model A because, "Fracking fluids and wastewater can be cause of normal tectonic plates," which reflects an error in conceptual understanding. We generally categorized participant explanations that discussed inaccurate links as erroneous, aside from clearly more advanced answers such as conscientious use of elimination-based logic. The second category, *descriptive evaluations*, represented weak and/or trivial written explanations. These weak explanations were generally from superficial evaluations between a line of evidence and a model. One student wrote that climate change evidence #1 strongly supports Model A because "they are both related to each other, they both talk about the same things." Although such evaluations were not necessarily inaccurate, they reflect little thinking and reasoning about the epistemic quality of the connection.

The third category, *relational evaluations*, represented correct links with somewhat deeper understanding, but participants' written explanations failed to differentiate between lines of evidence and explanatory models. For example, one student explained that fracking evidence #4 strongly supports Model A because "[the evidence] says that around fracking sites there are more earthquakes, therefore causing earthquakes." In this case, the student displays conceptual understanding about the evidence. However, the written explanation provides little insight into the epistemic level of quality applied in connecting the line of evidence to the explanation. The fourth and final category, *critical evaluations*, represented the greatest level of explanation development. Within this category, participants demonstrated an understanding of the scientific concepts and were able to critique the



Please work on this part individually after you complete your diagram. Now that you have completed the diagram, reconsider the plausibility of Models A and B.

Circle the plausibility of each model. [Make two circles, one for each model.]

	Greatly implausible (or even impossible)									Highly plausible
Model A	1	2	3	4	5	6	7	8	9	10
Model B	1	2	3	4	5	6	7	8	9	10

Did the plausibility of Model A and/or Model B change after you completed the diagram? Yes or No [Circle One]

[Note: you may have to look at your previous ratings if you do not remember what they were. Ask your teacher for assistance.]

Which arrows changed your plausibility judgments about the models? If your plausibility judgment did not change, which arrows supported your original plausibility judgments? Use the following steps to provide two explanations for why your plausibility judgments did or did not change.

A. Write the number of the evidence you are writing about. [Note: it is okay to include more than one evidence.]
B. Circle the appropriate word (**strongly supports | supports | contradicts | has nothing to do with**).
C. Write which model you are writing about. [Note: it is okay to include both models.]
D. Then write your reason.

1. Evidence # 1 **strongly supports | supports | contradicts | has nothing to do with Model** B **because:**
Fracking fluids and wastewater can be the cause of normal tectonic plates.

2. Evidence # 2 **strongly supports | supports | contradicts | has nothing to do with Model** B **because:**
The tectonic plates can maybe be the reason it was 11 times higher than a 30 year average.

Fig. 4. A student example of the fracking explanation tasks.

model-evidence links using scientific reasoning and an accurate representation of the role evidence plays in judging model validity. With these types of responses, students also demonstrated an explicit understanding of the epistemic quality of their connection between a line of evidence and an explanation. One student wrote that climate change evidence #2 contradicts model B because, "the evidence talks about how the Sun's energy is decreasing but model B is stating how the Sun's energy is increasing." Although the student does not specifically make a claim about the plausibility of the explanation in this response, the student is explicitly addressing the strength of the connection between the line of evidence and the explanation in way that evaluates the link's epistemic quality.

These four categories served as distinct levels of evaluation for numerically scoring each explanation (1 = erroneous, 2 = descriptive, 3 = relational, 4 = critical), allowing us to consider participants' written explanations quantitatively. The third and fourth authors independently scored each participant's explanations using these four categories and Lombardi, Brandt et al.'s (2016) rubric as a guide. Initial rater scores were at a very good level of agreement (interclass coefficient, *ICC* = 0.841). The raters met to reconcile discrepancies in scoring and came to unanimous agreement on explanation task scores. We used these agreed upon scores in the subsequent analysis.

2.2.3. Judgments of model plausibility

For each of the four MEL and MET activities, students recorded their plausibility judgments at pre and post instruction for each explanatory model they were shown. Students gauged the plausibility of each model using a 1–10 scale (1 = greatly implausible and 10 = highly plausible), based on earlier measures used by Lombardi et al. (2013), and Lombardi, Danielson et al. (2016). Specifically, Lombardi and others developed this 10-point scale based on previous plausibility instrumentation developed by cognitive scientists (see, for example, Connell & Keane, 2006). For the MEL and MET, we calculated plausibility scores as ratings for the scientific model minus ratings for the alternative model (Table 1). A positive score indicated that a participant judged the plausibility of the scientific model as greater than the alternative model, a negative rating indicated belief that the alternative model was more plausible, and a value of zero indicated belief that both models were equally plausible. For the Mono-MEL, we calculated the plausibility scores as the ratings for the scientific model minus the median rating for the class. In this case, a positive score indicated that a participant judged the model as more plausible than the class median and a negative score indicated that a participant judged the model as less plausible than the median. For the present study, the reliability of all plausibility scores exceeded the threshold considered minimally acceptable through previous meta-analysis of behavioral research studies (Peterson, 1994), with Guttman's $\lambda_2 = 0.68$,[2] and therefore, we decided to use these scores in our analysis.[3]

2.2.4. Knowledge

We created short, 5-item knowledge instruments for each topic (climate change, fracking, wetlands, and the Moon), which participants completed both prior to and just after engaging in a specific MEL activity. Per the methods used in our earlier MEL studies (see, for example, Lombardi et al., 2013), students rated each item on a 5-point Likert scale (1 = strongly disagree and 5 = strongly agree) indicating how closely they believed scientists would agree with the statement. In this way, answers reflected students' understandings about the related

[2] Guttman (1945) proposed some different measures to provide lower bound estimates for instrument reliability. Guttman based each measure on slightly different assumptions. For example, he based λ_3 on the restrictive assumption that individuals differ from each other in their true scores but each person has the same true score on each test (i.e., all covariances between items are equal). This is the essentially the same assumption that guided the development of Cronbach's α, and therefore, it is no surprise that $\alpha = \lambda_3$. The restrictive assumption of equal covariance is not part of the λ_2 calculation, and therefore, in virtually all classroom measurement situations, λ_2 is a more appropriate reliability measure than α (Woodruff & Wu, 2012).

[3] We acknowledge that some behavioral researchers consider 0.7 to be the cutoff threshold for reliable instrumentation because lower reliability tends to attenuate results due to higher signal to noise ratios. However, attenuation would be most pronounced at the ends of the sample distribution (Osterlind, 2010). As such, lower reliability most likely dampens differences in distribution samples (i.e., in the present study, lower reliability would dampen pre to post instruction differences; Carmines & Zeller, 1979). Therefore, it is extremely unlikely that lower reliability instruments would result in a Type II experimental error.

scientific processes rather than their personal beliefs or opinions on the topic. We developed these statements from information on which there is clear scientific consensus. We created these short forms from longer instruments after pilot study feedback revealed that teachers were spending too much instructional time on survey administration and students were viewing these longer instruments as unit tests. At least one question addressed each evidence statement in these short forms and our project's advisory panel verified the face and content validity of our items. For the present study, reliability of knowledge scores exceeded the threshold considered minimally acceptable through previous meta-analysis of behavioral research studies (Peterson, 1994), with Guttman's $\lambda_2 = 0.63$.

2.3. Procedures

During the summer prior to this quasi-experimental study, Ms. Rodgers and Ms. Williams participated in a three-day professional development workshop with the project team. Because Ms. Rodgers and Ms. Williams were completely new to the project, the master teachers who had participated in the first and second year pilot studies helped during the training. The workshops focused on practicing the MEL and comparison activities, going over the content and pedagogical strategies for effective classroom implementation, and planning for the upcoming year's implementation. To maintain some uniformity in instruction, Ms. Rodgers and Ms. Williams agreed to introduce each activity at the beginning of a unit prior to any instruction about the topic. The teachers also agreed to follow the lesson plans, as specified. In other words, the teachers presented the activities using the instructions present on the student materials. Class discussions that arose only centered on elaboration and clarification of these instructions. Because of the nature of the quasi-experimental design, each teacher agreed to teach the MEL to one class and a comparison activity (MET or Mono-MEL) to another class on the same days. We randomly assigned one class to the MEL and one to the comparison activity, and also randomly assigned Ms. Williams the MET as her comparison activity and Ms. Rodgers the Mono-MEL as her comparison activity.

Students completed all of the activities over the course of a single school year, which included the full activities and measures before and after all four topics were covered. A breakdown of activities completed by students and the order in which they were implemented for each MEL is provided in Fig. 5. Near the beginning of the year, prior to completing any activity, students performed the "plausibility ranking task" as an introduction to the ideas of plausibility and critical evaluation. This task asked students to rank the importance of different types of evidence for determining the plausibility of an explanatory model. These four types of evidence were the same as the links that students later indicated on the activities: evidence that supports the model, strongly supports it, contradicts it, or has nothing to do with it. After ranking the importance of each from 1 to 4, they read a small passage on falsifiability that states scientific ideas cannot be proven but are rather disproven through opposing evidence and were then asked to rank the types of evidence again. This provided an introduction to the idea of plausibility for students and an initial look at the ways in which they might evaluate connections between scientific evidence and explanations. Teachers had the option of repeating or discussing this activity as a review prior to covering the second topic if they felt it was needed.

Students completed each activity at the beginning of an instructional unit related to the topic (e.g., the climate change activity was conducted at the start of a unit on climate and weather, prior to any other instruction on the topic). For a given activity, students began by completing the knowledge test, if needed the plausibility ranking task described above, and model plausibility ratings for each explanatory model on that topic. At this time, teachers also engaged the class in an unscripted short discussion about the model(s) and the idea of plausibility, to clarify misunderstandings and address general questions about the topic. Students then began either the MEL diagram, the MET table, or the Mono-MEL diagram depending on which their class was randomly assigned. Students were able to read the evidence texts and complete the diagram or table in groups. They then worked individually to write up the explanation task. Each activity ended with the second iteration of the model plausibility ratings and knowledge test for that topic. Upon completion of this sequence, teachers moved on to teaching their regular instructional unit.

Each activity took place over about two regular class periods (~90 min total), with no appreciable difference in instructional time between the MEL, MET or Mono-MEL. The teachers implemented the activities during regular class time concurrently with their own planned curricula and at times when the topic of each MEL corresponded to scheduled lessons. As a result, the timing and order in which students completed the MEL was different based on the teachers' discretion. Ms. Rodgers did the wetlands activities in September, Moon activities in December, fracking activities in March, and climate activities in April. Ms. Williams did the climate activities in September, the Moon activities in January, the fracking activities in February, and the wetlands activities in April.

3. Results

We present the results in two sections. The first section discusses Research Question 1 and represents a relatively fine grained comparative analysis about the effectiveness of the instructional treatments (MEL, MET, and Mono-MEL) that we designed based on a theoretical position about learning controversial and complex science topics in a classroom setting. The second section discusses Research Question 2 and represents a somewhat broader analysis that examines relationships posited by this theoretical position.

3.1. Research question 1: Instructional treatment comparison

3.1.1. Preliminary analyses

We conducted repeated measures analyses of variance (ANOVAs) to compare the different instructional treatments (i.e., the MEL, MET, and Mono-MEL), with one ANOVA comparing pre and post instructional plausibility scores for each topic and one ANOVA comparing pre and post instructional knowledge scores. Prior to conducting these analyses,

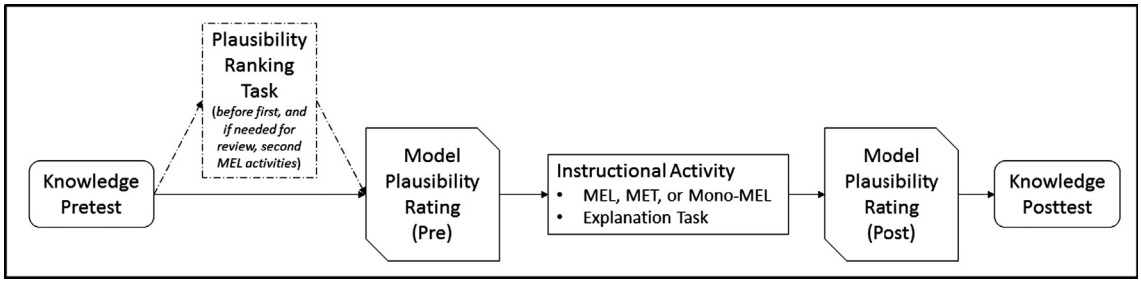


Fig. 5. The sequence of tasks for a given topic (i.e., climate change, fracking, wetlands, or Moon).

we conducted a preliminary check using repeated measures ANOVAs to compare both plausibility and knowledge scores, with classroom/teacher as the between-subjects variable, time (pre and post) as the within-subjects variable, and plausibility and knowledge scores the dependent variables, respectively. We only ran these repeated measures ANOVAs using MEL scores because each teacher tested a different comparison activity. The repeated measures ANOVAs revealed no significant interactions between the two classrooms/teachers and time, for both plausibility scores, with Wilks' $\lambda = 0.981$, $F(1, 28) = 0.530$, $p = .47$, and knowledge scores, with Wilks' $\lambda = 0.966$, $F(1, 28) = 0.981$, $p = .33$. Follow up simple effects analyses specifically showed no significant difference in classrooms/teachers at pre and post instruction for both plausibility and knowledge scores, with all p-values > .36. We also screened the data to ascertain alignment with assumptions inherent in ordinary least squares analyses (OLS; e.g., ANOVA) about the normality and linearity of the sample, as well as assumptions about the equality of the homogeneity of variance-covariance matrices. All of the variables had skewness and kurtosis of absolute value less than or equal to 1, which some researchers use as general rule of thumb to indicate normality of the sample distribution (Nussbaum, 2014). Our examination of scatterplots for pair combinations of the measured variables also did not reveal any concerns with linearity. Finally, variance-covariance matrices were equivalent for both plausibility (Box's $M = 13.9$, $p = .040$) and knowledge (Box's $M = 11.5$, $p = .091$).

3.1.2. Plausibility judgments

We conducted a repeated measures ANOVA to compare plausibility scores, where instructional treatment was the between-subjects variable, time (pre and post) was the within-subjects variable, and plausibility judgment score was the dependent variable (see Fig. 6 for plausibility scores at pre and post instruction by treatment). The repeated measures ANOVA indicated a significant interaction between treatment and time for plausibility, Wilks' $\lambda = 0.843$, $F(2, 61) = 5.67$, $p = .006$, with a medium effect size ($\eta^2 = 0.157$). We conducted a simple effects analysis, with Bonferroni adjustment for multiple comparisons, to analyze differences in scores at both pre and post instruction, as well as changes in scores from pre to post instruction for each treatment condition (see Fig. 6). The simple effects analysis revealed no significant difference between MEL, MET, and Mono-MEL plausibility scores at pre instruction (all p-values > .38), However, post instruction MEL plausibility scores ($M = 1.76$, $SD = 1.42$) were significantly greater than both post instruction MET scores ($M = 0.74$, $SD = 1.21$), with $p = .032$; and post instruction Mono-MEL scores ($M = 0.00$, $SD = 0.901$), with $p < .001$. The simple effects analysis also revealed significant increases in MEL and MET plausibility scores from pre instruction (MEL, $M = 0.59$, $SD = 1.42$; and MET, $M = -0.03$, $SD = 1.44$) to post instruction, with $F(1, 61) = 20.2$, $p < .001$, $\eta^2 = 0.249$ (large effect size) for the MEL, and $F(1, 61) = 4.93$, $p = .03$, $\eta^2 = 0.075$ (small effect size) for the MET. However, there was not a significant change in Mono-MET scores from pre instruction ($M = 0.28$, $SD = 1.00$) to post instruction ($M = 0.00$, $SD = 0.901$), with $p = .42$.

3.1.3. Knowledge

We conducted a repeated measures ANOVA to compare knowledge scores, where instructional treatment was the between-subjects variable, time (pre and post) was the within-subjects variable, and knowledge score was the dependent variable (see Fig. 7 for knowledge scores at pre and post instruction by treatment). The repeated measures ANOVA revealed a significant interaction between treatment and time for knowledge, Wilks' $\lambda = 0.893$, $F(2, 61) = 3.67$, $p = .03$, with a medium effect size ($\eta^2 = 0.107$). We conducted a simple effects analysis, with Bonferroni adjustment for multiple comparisons, to analyze differences in scores at both pre and post instruction, as well as changes in scores from pre to post instruction for each treatment condition (see Fig. 7). The simple effects analysis revealed no significant difference between MEL, MET, and Mono-MEL knowledge scores at pre instruction (all p-values > .36) and post instruction (all p-values > .40). However, there were significant increases in MEL scores from pre to post instruction (pre, $M = 69.2$, $SD = 6.36$, and post, $M = 73.9$, $SD = 5.56$), with $F(1, 61) = 14.5$, $p < .001$, $\eta^2 = 0.192$ (medium effect size). There were also significant increases in MET knowledge scores from pre to post instruction (pre, $M = 69.4$, $SD = 5.84$, and post, $M = 73.1.4$, $SD = 7.21$), with $F(1, 61) = 4.96$, $p = .030$, $\eta^2 = 0.075$ (small effect size). However, there was not a significant change in Mono-MET scores from pre instruction ($M = 71.9$, $SD = 3.31$) to post instruction ($M = 71.1$, $SD = 6.11$), with $p = .64$.

The increase in MEL knowledge scores is about 6.8%, with increases in MET knowledge scores of about 5.3% (i.e., about a half a letter grade

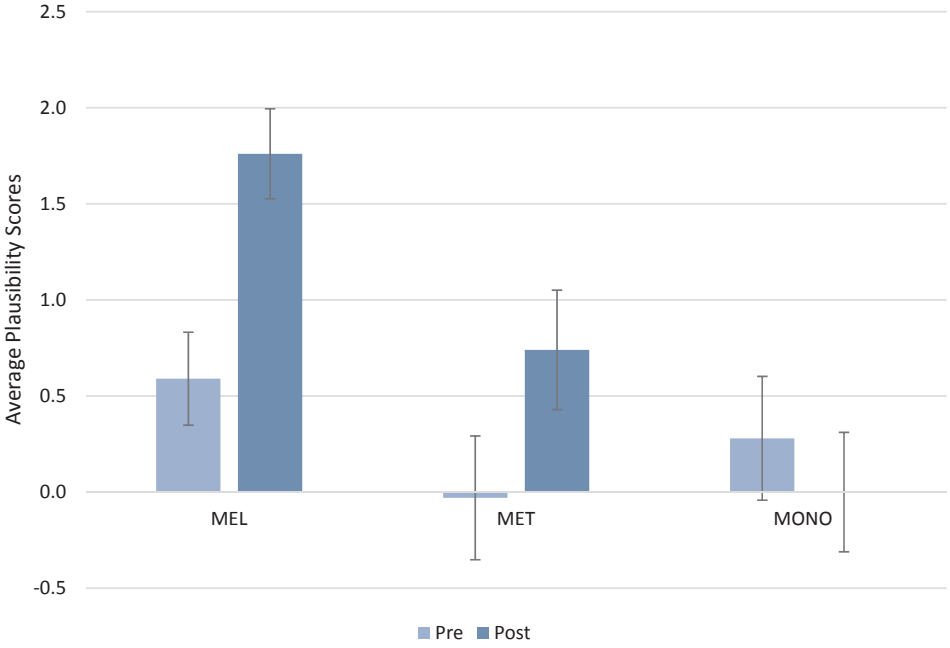


Fig. 6. Average pre and post instructional plausibility scores by treatment activity. The possible score range was from −9 to +9. Bars on each column indicate ± 1 standard error.

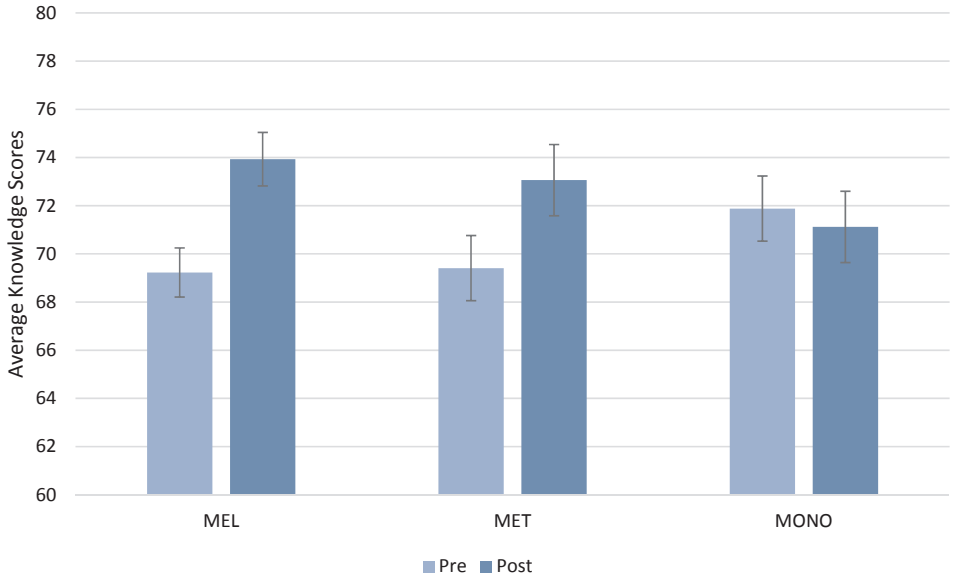


Fig. 7. Pre and post instructional knowledge scores by treatment activity. The possible score range was from 20 to 100. Bars on each column indicate ± 1 standard error.

increase). Within the context of the present study (i.e., situated within authentic classroom instructional settings), these increases are robust. Each of the activities lasted only about two class periods of traditional instruction (eight class days total). When considering that a year of instruction is approximately 180 days, such increases may have a strong practical significance for classroom instruction.

3.2. Research question 2: Relations between evaluation, plausibility, and knowledge

Fig. 8 shows a model of the relations between evaluation (as measured by the activities' explanation tasks), plausibility (at both pre and post instruction), and knowledge (at both pre and post instruction), and also includes antecedent variables that may influence these relations, including the instructional treatment (MEL, MET, and Mono-MEL, coded as 3, 2, and 1 as ordered categories, respectively), teacher (Ms. Williams and Ms. Rodgers, coded categorically as 1 and 2, respectively), and socioeconomic status (SES; we used participants' application for federal free or reduced-price student meals—1 for an application and 0 for no application—as a surrogate for SES, which is a common practice in educational research; Sirin, 2005). Including both teacher and SES as variables examines the potential effects from

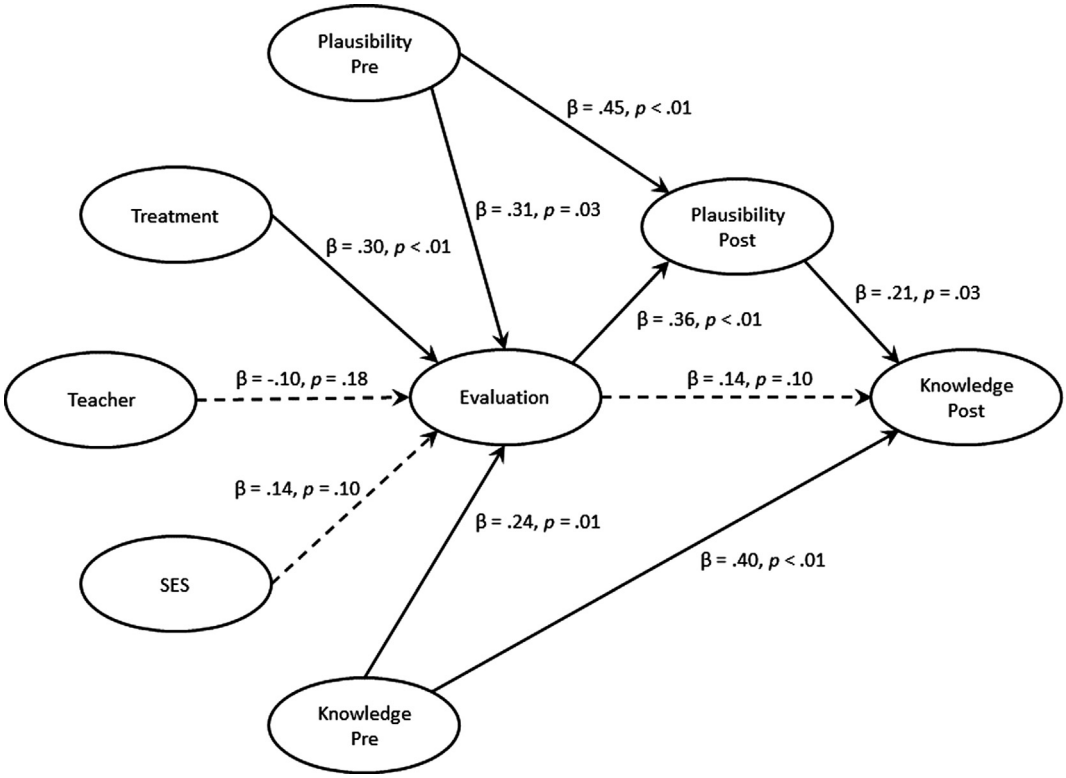


Fig. 8. Model of the relations between study variables. Solid lines indicate significant pathways and dashed lines indicated non-significant pathways.

different classrooms, different schools, and different regional settings. We adopted a null hypothesis (i.e., there would be no significant relations with these effects) because of disparity in results in the extant critical thinking and reasoning literature (see, for example, Abrami et al., 2008). The relational paths between evaluation, plausibility, and knowledge reflect Lombardi, Nussbaum et al. (2016) theoretical model and previous empirical research examining these relations (Lombardi, Danielson et al., 2016, Lombardi et al., 2018).

We used variance-based structural equation modeling (VB-SEM) to examine the relational paths shown in Fig. 8, and specifically used the Warp PLS v.4.0 statistical software (Kock, 2013). Unlike traditional, covariance-based structural equation modeling, which assumes that all data is metric/continuous and conforms to a normal distribution, VB-SEM uses the partial least-squares method, which is based on ranked data and is distribution-free. Use of ranked-based data allows for more statistical power without compromising or inflating the chance for Type I errors for a large range of sample sizes and variation of group sizes (Reinartz, Haenlein, & Henseler, 2009). VB-SEM and partial least-squares methods have been used increasingly in social science research (Esposito Vinzi, Chin, Henseler, & Wang, 2010), and are being used more frequently in educational research (see, for example, Hagger, Sultan, Hardcastle, & Chatzisarantis, 2015; Lombardi, Danielson et al., 2016).

We constructed the latent variables in our model (evaluation; pre and post instructional plausibility; and pre and post instructional knowledge) using scores for all four topics (climate change, fracking, wetlands, and Moon), respectively. Table 2 shows the first order bivariate correlations, means, and standard deviations for each variable.

We used several fit and quality indices to gauge the validity of our variance-based structural equation model. These indices include overall goodness-of-fit (GoF), average path coefficient (APC), average coefficient of determination across the model (average R^2 or ARS), average variance inflation factor for model parameters (AVIF), and average full collinearity VIF (AFVIF). Tenenhaus, Amato, and Esposito Vinzi (2004) proposed that researchers use GoF as a criterion for the overall model prediction performance based on both the measurement and the structural model. A model has a large explanatory power when GoF is greater than 0.36, with unacceptable explanatory power when GoF is less than 0.1 (Wetzels, Odekerken-Schroder, & van Oppen, 2009). Both APC and ARS provide further information about model adequacy and together gauge the predictive and explanatory power of the model (analogous to total variance explained). APC and ARS should have values that are statistically significant, with *p*-values less than 0.05 generally considered acceptable (Hagger et al., 2015). High AVIF and AFVIF values indicate a potentially large degree of collinearity (i.e., redundancy of variables; Tabachnick & Fidell, 2007) is present in the model. Values of AVIF and AFVIF should generally be below 3.3 (Kock & Lynn, 2012) to ensure that variables are not redundant. Finally, nonlinear bivariate causality direction ratio (NLBCDR) indicates the percentage of model paths where the hypothesized direction is supported, with "acceptable values of NLBCDR…equal to or greater than 0.7, meaning that in at least 70 percent of path-related instances in a model the support for the reversed hypothesized direction of causality is weak or less" (Kock, 2013, p. 53). For the present study, the overall fit and quality of model was excellent, with GoF = 0.437 (large explanatory power; Tenenhaus, Esposito Vinzi, Chatelin, & Lauro, 2005); APC = 0.265, $p < .001$; ARS = 0.330, $p < .001$; AVIF = 1.12; AFVIF = 1.46; and NLBCDR = 1.0 (100% support for the hypothesized path directions; e.g., evaluation to plausibility to knowledge; Fig. 8).

We included the model's standardized path values in Fig. 8. We chose standardized values because this allows the reader to compare differences of magnitude between predictors with different scales. Of particular note are the direct paths between evaluation and post instructional plausibility, and post instructional plausibility and knowledge, both of which have significant standardized path values. In contrast, the direct path between evaluation and post instructional knowledge is not significant. Thus, the SEM indicates that the path relating evaluation, post instructional plausibility and knowledge supports Lombardi, Nussbaum et al.'s (2016) theoretical model. Furthermore, the contribution of evaluation to post instructional plausibility is just slightly less than the contribution of pre instructional plausibility ($\beta = 0.36$ vs. $\beta = 0.45$), providing some support that instruction promoted plausibility reappraisal. Finally, the two other non-significant paths (teacher to evaluation and SES to evaluation) are notable and show that neither classroom, district, nor regional differences has an influence on the instructional treatment. However, the path between instructional treatment and evaluation was significant. Because we coded the MEL as 3, MET as 2, and Mono-MEL as 1, the positive standardized path value ($\beta = 0.30$) shows that MEL had a greater influence on evaluations than either the MET or Mono-MEL.

We also calculated total effect sizes for each variable in relation to evaluation, post instructional plausibility, and post instructional knowledge, but only for the significant pathways. Effect sizes are useful to help ascertain strength and meaning of the variable relations (i.e., practical significance of one variable on another). We specifically used Cohen's (1988) f^2, which researchers use to gauge one variable's effect size within the context of the related variables. As shown in Table 3, both the effects of plausibility and knowledge at pre instruction are large on plausibility and knowledge at post instruction, respectively. This is no surprise given the importance of background state on learning (see, for example, McNamara & Kintsch, 1996). A more novel finding is the large effect size of evaluation on post instructional plausibility, which shows that greater levels of evaluation led to greater plausibility of the scientific alternative and suggests the potential influence of plausibility reappraisal. Another noteworthy result is that all of relational effects were near to or greater than the medium size threshold of 0.15, including the effect of evaluation on both post instructional plausibility and knowledge.

Table 2
Bivariate correlations, means, and standard deviations for evaluation, plausibility, and knowledge scores.

	1	2	3	4	5
1. Evaluation	–				
2. Plausibility (pre instruction)	0.246*	–			
3. Plausibility (post instruction)	0.460**	0.540**	–		
4. Knowledge (pre instruction)	0.096	0.257**	0.163	–	
5. Knowledge (post instruction)	0.292*	0.334**	0.355**	0.385**	–
M	2.16	0.34	1.02	70.0	73.0
SD	0.289	1.33	1.47	5.61	6.11

Note. For ease of interpretation, evaluation and plausibility scores are averages for each MEL activity (total scores divided by 4), where evaluation scores had a possible range from 1 to 4 and plausibility scores had a possible range of −9 to +9. However, knowledge scores represent totals, with a possible range of 20–100.

* $p < .05$.

** $p < .01$.

4. Discussion

The results revealed that MEL diagrams promoted high school students' evaluations of the connections between lines of evidence and alternative explanations about various Earth science phenomena, including causes of current climate change, relations between fracking and earthquakes, use of wetlands, and formation of Earth's Moon. Deeper evaluations, in turn, helped students to reappraise their plausibility judgments toward more scientific explanations and construct more scientifically accurate knowledge. The direct pathway between evaluation and knowledge was non-significant, suggesting that students may need to more explicitly consider their appraisals and reappraisals of plausibility for deeper understanding per Lombardi, Nussbaum

Table 3
Total effect sizes, as Cohen's (1988) f^2 values, for the significant relational paths.

End of path variable	Beginning or along path variable				
	Plausibility Pre	Knowledge Pre	Treatment	Evaluation	Plausibility Post
Evaluation	0.311	0.243	0.298	–	–
Plausibility Post	0.561	0.087	0.107	0.360	–
Knowledge Post	0.163	0.450	0.065	0.219	0.210

Note. $f^2 \geq 0.02$ = small effect size, $f^2 \geq 0.15$ = medium effect size; $f^2 \geq 0.35$ = large effect size.

et al.'s (2016) theoretical position. Such evaluations and judgments likely do not always factor into science learning, but they may be especially relevant for topics with a large plausibility gap (e.g., climate change and fracking).

The analysis associated with Research Question 1 showed that both the MELs and METs were generally more effective at promoting plausibility reappraisal and knowledge construction than the Mono-MELs, both with medium to large effect sizes for the MEL and small effect sizes for the MET. For the analysis associated with Research Question 2, effect sizes of individual causal pathways were medium to large, suggesting meaningful and practical relations between students' evaluations, plausibility reappraisal, and knowledge construction. This specific finding supports the notion that the process of evaluation and increasing plausibility of scientific explanations can support deeper understanding about certain topics (Lombardi, Nussbaum et al., 2016). Because of the context of this study (i.e., situated within high school classrooms, with relatively a brief amount of instructional time), these effect sizes suggest some relevance for classroom practice. Specifically, increases in knowledge scores represent about a half a letter grade increase, which is meaningful given that total dosage of instruction was only eight class days (i.e., about 5% of the total instructional dosage in a typical 180-day school year). Such increases may have a strong practical significance for classroom instruction. Both the MEL and MET scaffolds facilitate students' evaluation between lines of evidence and two alternative explanations, as opposed to the Mono-MEL that only included one alternative. Therefore, designing instructional scaffolds that more closely reflect scientific and critical evaluations (i.e., evaluations involving more than one alternative explanations) has implications for classroom science learning (see, for example, Azevedo & Hadwin, 2005; Pea, 2004; Quintana et al., 2004; Van Merriënboer, Kirschner, & Kester, 2003). This is a particularly relevant point because teacher and SES (i.e., as a surrogate for local and regional differences in external factors that could impact learning) were not significantly related to students' evaluations, plausibility judgments, and knowledge construction. This suggests that collaborative classroom-based research can develop robust instructional scaffolds to facilitate classroom instruction that, in turn, helps all students think scientifically.

Our project and the present study are limited by the nature of the MEL activity, which is of relatively short duration (i.e., ~90 min of instructional time per topic). We purposefully designed the MEL to fit modularly within a longer two- to three-week instructional unit in order to maximize dissemination and potential usefulness to classroom teachers. However, preliminary results show that students who engage in all four MELs during the course of a school year have difficulty in transferring their critical evaluations to a more distal task. Of course, transfer of learning is challenging, and transfer of critical thinking and scientific reasoning may even be a greater challenge. Because of its relatively short duration, the MEL activity may not be optimal in facilitating students' conceptual agency (i.e., students' appropriation and modification of materials as conceptual resources for the purpose of successfully completing an authentic task; Pickering, 1995). This idea of conceptual agency reflects Kuhn's (2010) notion of meta-knowledge that students may acquire during argumentation activities. For effective transfer, students should internalize instructional scaffolds via conceptual agency and development of meta-knowledge to construct robust mental representations for application (Nussbaum & Asterhan, 2016). Therefore, revising the MEL and/or designing other scaffolds to facilitate potential increases in conceptual agency and meta-knowledge may be warranted to assist in transferring critical evaluations and plausibility reappraisals beyond the classroom environment.

We specifically designed the MEL to be a pencil and paper activity given the necessities of our partner schools. However, development of digital MELs that students can access and manipulate via modern electronic technologies (e.g., computers, tablets, smart phones) may increase potential usability, particularly for those schools that are actively using such technologies for instruction. We speculate that there are several differences in outcomes that may occur between virtual and pencil-and-paper versions of the MEL. First, teacher-student and student-student interactions may vary between the activities, specifically in the types and degree of engagement in the scientific practices (see, for example, Gobert, Baker, & Wixon, 2015). Second, with the virtual MELs, students might be able to access more information (e.g., via hyperlink clicks) than provided by the evidence texts (i.e., the one-page, hard copies of information about each line of evidence). It is possible that ease of clicking and availability of information may promote a different level of engagement with evidence texts compared to having hard copy pages available for the students (see, for example, Bråten, Strømsø, & Britt, 2009). Third, we speculate that there may be different qualities in the explanation tasks between the versions. In a pencil-and-paper MEL, students hand write their explanations, and in a virtual MEL, students would type their explanations, potentially resulting in memory encoding and learning differences (Mayer & Moreno, 1998). With the growing evidence base associated with the pencil and paper version, these speculations could provide promising avenues of research for those interested in learning with modern digital technology.

We acknowledge that reducing the length of the knowledge instruments resulted in lower score reliability, which in turn, warrants caution in interpreting the results. However, we stress that lower reliability generally means that results would be more attenuated at the ends of sample distributions. This would probably dampen the pre to post instructional differences revealed in the fine-grained analyses that we conducted to investigate Research Question 1. In other words, these differences, all with medium effect sizes, would most likely be stronger if we had used a knowledge instrument with more items.

4.1. Educational implications

Our research suggests that the MEL activities can help students to think more scientifically during the process of knowledge construction, but these scaffolds are not a panacea. For instance, these activities feature some scientific practices (e.g., engaging in critique and argument, analyzing lines of evidence, using models), but not others (e.g., making observations, planning investigations, and collecting data). However, the results from the present study and our previous studies consistently suggest that students should evaluate connections between lines of evidence and alternative explanations about phenomena, particularly those phenomena that may have a large plausibility gap (i.e., where lay people may judge scientific explanations to be relatively implausible). Although scientific hypotheses, models, and theories

undergo certain evaluative processes that increase their perceived truthfulness, teachers should not assume that students fully understand these processes and render the same judgments as the scientific community. Rather, for students to gain a deep understanding of science, instruction should encourage students to evaluate their own knowledge in light of scientific evidence and facilitate collaboration and critique during science learning (NRC, 2012). Scaffolds, such as the MEL, hold promise in this regard, particularly for complex and abstract concepts where students have the opportunity to consider competing explanations and use evidence to gauge the pros and cons of each. Students should also engage in constructive critique and evaluation throughout an instructional unit. For example, when collecting data students should reflect on the reliability of their measurements (as we have done in reporting the present study). Although researchers have encouraged such an instructional approach (see, for example, Bailin, 2002; Berland & Reiser, 2009; Duschl et al., 2007; Ford, 2015), to our knowledge, curriculum developers and teachers are not promoting the process of critical evaluation as an essential element of scientific thinking and knowledge.

4.2. Concluding remarks

We introduced the notion of scientific literacy right at the outset. Although some consider the term to be a "weasel word," we agree with Dillon (2016) "…that scientific literacy…will be part of the discourse of science education for a long time" (p. 271). To add to this discourse and operationalize scientific literacy, we have contextualized the phenomenon in terms of current science education reform efforts endeavoring to deepen students' knowledge construction through engagement in scientific practices and scientific thinking. Critical evaluation and plausibility reappraisal, specifically when considering the connections between lines of evidence and alternative explanations, are a synergistic process of scientific thinking and knowing (Lombardi, Nussbaum et al., 2016). Furthermore, a growing body of evidence, from our research and others (see, for example, Ranney, Munnich, & Lamprey, 2016), suggests that instructional scaffolds can help students think more critically, facilitate epistemic inferences and judgments, and deepen students' science knowledge in classroom settings. Our hope in conducting such research is to inform future instructional design through pragmatic application of learning theories, all which help to equip a more scientifically literate citizenry able to equitably solve local, regional, and global problems.

Acknowledgments

We wish to express sincere thanks to Ms. Rodgers and Ms. Williams, and their students, for inviting us into their classrooms and participating in the present study. We also wish to thank our master teachers and students who participated in our earlier pilot studies, as well as our project's advisory board: Dr. Clark A. Chinn, Rutgers University; Dr. Bruce E. Herbert, Texas A&M University; Dr. Kim A. Kastens; Lamont-Doherty Earth Observatory; and Dr. Gale M. Sinatra, University of Southern California. Finally, the National Science Foundation (NSF), under Grant No. DRL-1316057 and Grant No. DRL-1721041, supported this research. Any opinions, findings, conclusions, or recommendations expressed are those of the authors and do not necessarily reflect the NSF's views.

References

Abrami, P. C., Bernard, R. M., Borokhovski, E., Wade, A., Surkes, M. A., Tamim, R., & Zhang, D. (2008). Instructional interventions affecting critical thinking skills and dispositions: A stage 1 meta-analysis. *Review of Educational Research, 78*(4), 1102–1134. http://dx.doi.org/10.3102/0034654308326084.

Anderson, T., & Shattuck, J. (2012). Design-based research: A decade of progress in education research? *Educational Researcher, 41*(1), 16–25. http://dx.doi.org/10.3102/0013189X11428813.

Azevedo, R., & Hadwin, A. F. (2005). Scaffolding self-regulated learning and metacognition – Implications for the design of computer-based scaffolds. *Instructional Science, 33*, 367–379. http://dx.doi.org/10.1007/s11251-005-1272-9.

Bailin, S. (2002). Critical thinking and science education. *Science & Education, 11*(4), 361–375. http://dx.doi.org/10.1023/A:1016042608621.

Barab, S., & Squire, K. (2004). Design-based research: Putting a stake in the ground. *The Journal of the Learning Sciences, 13*(1), 1–14. http://dx.doi.org/10.1207/s15327809jls1301_1.

Berland, L. K., & Reiser, B. J. (2009). Making sense of argumentation and explanation. *Science Education, 93*(1), 26–55. http://dx.doi.org/10.1002/sce.20286.

Beyer, B. K. (1995). *Critical thinking.* Bloomington, IN: Phi Delta Kappa Educational Foundation.

Braaten, M., & Windschitl, M. (2011). Working toward a stronger conceptualization of scientific explanation for science education. *Science Education, 95*(4), 639–669. http://dx.doi.org/10.1002/sce.20449.

Bråten, I., Strømsø, H. I., & Britt, M. A. (2009). Trust matters: Examining the role of source evaluation in students' construction of meaning within and across multiple texts. *Reading Research Quarterly, 44*(1), 6–28. http://dx.doi.org/10.1598/RRQ.44.1.1.

Carmines, E. G., & Zeller, R. A. (1979). *Reliability and validity assessment.* Newbury Park, CA: Sage Publications.

Chi, M. T. H. (2005). Commonsense conceptions of emergent processes: Why some misconceptions are robust. *Journal of the Learning Sciences, 14*(2), 161–199. http://dx.doi.org/10.1207/s15327809jls1402_1.

Chinn, C., & Brewer, W. (1993). The role of anomalous data in knowledge acquisition: A theoretical framework and implications for science education. *Review of Educational Research, 63*, 1–49. http://dx.doi.org/10.3102/00346543063001001.

Chinn, C. A., & Brewer, W. (2001). Models of data: A theory of how people evaluate data. *Cognition and Instruction, 19*, 323–398. http://dx.doi.org/10.1207/S1532690XCI1903_3.

Chinn, C., & Buckland, L. (2012). Model-based instruction: Fostering change in evolutionary conceptions and epistemic practices. In K. S. Rosengren, E. M. Evans, S. Brem, & G. M. Sinatra (Eds.). *Evolution challenges: Integrating research and practice in teaching and learning about evolution* (pp. 211–232). New York, NY: Oxford University Press.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* Hillsdale, NJ: Lawrence Erlbaum.

Collins, A., & Michalski, R. (1989). The logic of plausible reasoning: A core theory. *Cognitive Science, 13*, 1–49. http://dx.doi.org/10.1207/s15516709cog1301_1.

Conant, J. B. (1951). *Science and common sense.* New Haven, CT: Yale University Press.

Connell, L., & Keane, M. T. (2006). A model of plausibility. *Cognitive Science, 30*, 95–120. http://dx.doi.org/10.1207/s15516709cog0000_53.

Dillon, J. (2016). *Towards a convergence between science and environmental education: The selected works of Justin Dillon.* New York, NY: Routledge, Taylor & Francis Group.

Dole, J. A., & Sinatra, G. M. (1998). Reconceptualizing change in the cognitive construction of knowledge. *Educational Psychologist, 33*, 109–128. http://dx.doi.org/10.1080/00461520.1998.9653294.

Driver, R., Leach, J., Millar, R., & Scott, P. (1996). *Young people's images of science.* Buckingham, England: Open University Press.

Duschl, R. (2008). Science education in three-part harmony: Balancing conceptual, epistemic, and social learning goals. *Review of Research in Education, 32*(1), 268–291. http://dx.doi.org/10.3102/0091732X07309371.

Duschl, R. A., Schweingruber, H. A., & Shouse, A. E. (Eds.). (2007). *Taking science to school: Learning and teaching science in grades K-8.* Washington, DC: National Academies Press.

Erduran, S., & Dagher, Z. R. (2014). *Reconceptualizing the nature of science for science education.* Dordrecht, Netherlands: Springer.

Esposito Vinzi, V., Chin, W. W., Henseler, J., & Wang, H. (2010). *Handbook of partial least squares: Concepts, methods, and application.* Berlin, Germany: Springer.

Ford, M. J. (2015). Educational implications of choosing "practice" to describe science in the Next Generation Science Standards. *Science Education, 99*(6), 1041–1048. http://dx.doi.org/10.1002/sce.21188.

Gijlers, H., & de Jong, T. (2009). Sharing and confronting propositions in collaborative inquiry learning. *Cognition and Instruction, 27*(3), 239–268. http://dx.doi.org/10.1080/07370000903014352.

Gobert, J. D., Baker, R. S., & Wixon, M. B. (2015). Operationalizing and detecting disengagement within online science microworlds. *Educational Psychologist, 50*(1), 43–57. http://dx.doi.org/10.1080/00461520.2014.999919.

Greene, J. A., Hutchison, L. A., Costa, L. J., & Crompton, H. (2012). Investigating how college students' task definitions and plans relate to self-regulated learning processing and understanding of a complex science topic. *Contemporary Educational Psychology, 37*(4), 307–320. http://dx.doi.org/10.1016/j.cedpsych.2012.02.002.

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika, 10*(4), 255–282. http://dx.doi.org/10.1007/BF02288892.

Hagger, M. S., Sultan, S., Hardcastle, S. J., & Chatzisarantis, N. L. (2015). Perceived autonomy support and autonomous motivation toward mathematics activities in educational and out-of-school contexts is related to mathematics homework behavior and attainment. *Contemporary Educational Psychology, 41*, 111–123. http://dx.doi.org/10.1016/j.cedpsych.2014.12.002.

Hogan, K., & Maglienti, M. (2001). Comparing the epistemological underpinnings of students' and scientists' reasoning about conclusions. *Journal of Research in Science Teaching, 38*(6), 663–687. http://dx.doi.org/10.1002/tea.1025.

Kapon, S., & diSessa, A. A. (2012). Reasoning through instructional analogies. *Cognition and Instruction, 30*, 261–310. http://dx.doi.org/10.1080/07370008.2012.689385.

Klopfer, L. E. (1969). The teaching of science and the history of science. *Journal of Research in Science Teaching, 6*(1), 87–95. http://dx.doi.org/10.1002/tea.3660060116.

Kock, N. (2013). *WarpPLS 4.0 user manual.* Laredo, TX: ScriptWarp Systems.
Kock, N., & Lynn, G. S. (2012). Lateral collinearity and misleading results in variance-based SEM: An illustration and recommendations. *Journal of the Association for Information Systems, 13*(7), 546–580.
Kuhn, D. (1999). A developmental model of critical thinking. *Educational Researcher, 28*(2), 16–46. http://dx.doi.org/10.3102/0013189X028002016.
Kuhn, D. (2010). Teaching and learning science as argument. *Science Education, 94*(5), 810–824. http://dx.doi.org/10.1002/sce.20395.
Kuhn, D., & Pearsall, S. (2000). Developmental origins of scientific thinking. *Journal of Cognition and Development, 1*, 113–129. http://dx.doi.org/10.1207/S15327647JCD0101N_11.
Kyza, E. A. (2009). Middle-school students' reasoning about alternative hypotheses in a scaffolded, software-based inquiry investigation. *Cognition and Instruction, 27*(4), 277–311. http://dx.doi.org/10.1080/07370000903221718.
LaBerge, D., & Samuels, S. J. (1974). Toward a theory of automatic information processing in reading. *Cognitive Psychology, 6*(2), 293–323. http://dx.doi.org/10.1016/0010-0285(74)90015-2.
Li, M., Murphy, P. K., Wang, J., Mason, L. H., Firetto, C. M., Wei, L., & Chung, K. S. (2016). Promoting reading comprehension and critical–analytic thinking: A comparison of three approaches with fourth and fifth graders. *Contemporary Educational Psychology, 46*, 101–115. http://dx.doi.org/10.1016/j.cedpsych.2016.05.002.
Lombardi, D., Bickel, E. S., Bailey, J. M., & Burrell, S. (2018). High school students' evaluations, plausibility (re) appraisals, and knowledge about topics in Earth science. *Science Education, 102*(1), 153–177. http://dx.doi.org/10.1002/sce.21315.
Lombardi, D., Brandt, C. B., Bickel, E. S., & Burg, C. (2016). Students' evaluations about climate change. *International Journal of Science Education, 38*(8), 1392–1414. http://dx.doi.org/10.1080/09500693.2016.1193912.
Lombardi, D., Danielson, R. W., & Young, N. (2016). A plausible connection: Models examining the relations between evaluation, plausibility, and the refutation text effect. *Learning and Instruction, 44*, 74–86. http://dx.doi.org/10.1016/j.learninstruc.2016.03.003.
Lombardi, D., Nussbaum, E. M., & Sinatra, G. M. (2016). Plausibility judgments in conceptual change and epistemic cognition. *Educational Psychologist, 51*(1), 35–56. http://dx.doi.org/10.1080/00461520.2015.1113134.
Lombardi, D., & Sinatra, G. M. (2012). College students' perceptions about the plausibility of human-induced climate change. *Research in Science Education, 42*, 201–217. http://dx.doi.org/10.1007/s11165-010-9196-z.
Lombardi, D., & Sinatra, G. M. (2013). Emotions about teaching about human-induced climate change. *International Journal of Science Education, 35*, 167–191. http://dx.doi.org/10.1080/09500693.2012.738372.
Lombardi, D., Sinatra, G. M., & Nussbaum, E. M. (2013). Plausibility reappraisals and shifts in middle school students' climate change conceptions. *Learning and Instruction, 27*, 50–62. http://dx.doi.org/10.1016/j.learninstruc.2013.03.001.
Mason, L., Ariasi, N., & Boldrin, A. (2011). Epistemic beliefs in action: Spontaneous reflections about knowledge and knowing during online information searching and their influence on learning. *Learning and Instruction, 21*(1), 137–151. http://dx.doi.org/10.1016/j.learninstruc.2010.01.001.
Mayer, R. E., & Moreno, R. (1998). A split-attention effect in multimedia learning: Evidence for dual processing systems in working memory. *Journal of Educational Psychology, 90*(2), 312. http://dx.doi.org/10.1037/0022-0663.90.2.312.
McNamara, D. S., & Kintsch, W. (1996). Learning from texts: Effects of prior knowledge and text coherence. *Discourse Processes, 22*(3), 247–288. http://dx.doi.org/10.1080/01638539609544975.
McNeill, K. L., Lizotte, D. J., Krajcik, J., & Marx, R. W. (2006). Supporting students' construction of scientific explanations by fading scaffolds in instructional materials. *Journal of the Learning Sciences, 15*(2), 153–191. http://dx.doi.org/10.1207/s15327809jls1502_1.
Metz, K. E. (2004). Children's understanding of scientific inquiry: Their conceptualization of uncertainty in investigations of their own design. *Cognition and Instruction, 22*(2), 219–290. http://dx.doi.org/10.1207/s1532690xci2202_3.
National Research Council (1996). *National science education standards.* Washington, DC: National Academy Press.
National Research Council (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas.* Washington, DC: National Academies Press.
NGSS Lead States (2013). *Next generation science standards: For states by states.* Washington, DC: The National Academies Press.
Nussbaum, E. M. (2008). Collaborative discourse, argumentation, and learning: Preface and literature review. *Contemporary Educational Psychology, 33*, 345–359. http://dx.doi.org/10.1016/j.cedpsych.2008.06.001.
Nussbaum, E. M. (2011). Argumentation, dialogue theory, and probability modeling: Alternative frameworks for argumentation research in education. *Educational Psychologist, 46*, 84–106. http://dx.doi.org/10.1080/00461520.2011.558816.
Nussbaum, E. M. (2014). *Categorical and nonparametric data analysis.* New York, NY: Routledge.
Nussbaum, E. M., & Asterhan, C. S. (2016). The psychology of far transfer from classroom argumentation. In F. Paglieri (Ed.). *The psychology of argument: Cognitive approaches to argumentation and persuasion.* London, UK: College Publications, Studies in Logic and Argumentation Series.
Nussbaum, E. M., & Edwards, O. V. (2011). Critical questions and argument stratagems: A framework for enhancing and analyzing students' reasoning practices. *Journal of the Learning Sciences, 20*(3), 443–488. http://dx.doi.org/10.1080/10508406.2011.564567.
Osterlind, S. J. (2010). *Modern measurement: Theory, principles, and applications of mental appraisal.* Boston, MA: Allyn & Bacon.
Pea, R. D. (2004). The social and technological dimensions of scaffolding and related theoretical concepts for learning, education, and human activity. *Journal of the Learning Sciences, 13*(3), 423–451. http://dx.doi.org/10.1207/s15327809jls1303_6.
Peterson, R. A. (1994). A meta-analysis of Cronbach's coefficient alpha. *Journal of Consumer Research, 21*(2), 381–391. http://dx.doi.org/10.1086/209405.
Pickering, A. (1995). *The mangle of practice: Time, agency, and science.* Chicago: University of Chicago Press.
Pintrich, P. R., Marx, R. W., & Boyle, R. B. (1993). Beyond cold conceptual change: The role of motivational beliefs and classroom contextual factors in the process of conceptual change. *Review of Educational Research, 63*, 167–199. http://dx.doi.org/10.3102/00346543063002167.
Posner, G. J., Strike, K. A., Hewson, P. W., & Gertzog, W. A. (1982). Accommodation of a scientific conception: Toward a theory of conceptual change. *Science Education, 66*(2), 211–227. http://dx.doi.org/10.1002/sce.3730660207.
Quintana, C., Reiser, B. J., Davis, E. A., Krajcik, J., Fretz, E., Duncan, R. G., ... Soloway, E. (2004). A scaffolding design framework for software to support science inquiry. *Journal of the Learning Sciences, 13*(3), 337–386. http://dx.doi.org/10.1207/s15327809jls1303_4.
Ranney, M. A., Munnich, E. L., & Lamprey, L. N. (2016). Chapter 4-Increased wisdom from the ashes of ignorance and surprise: Numerically-driven inferencing, global warming, and other exemplar realms. *Psychology of Learning and Motivation, 65*, 129–182. http://dx.doi.org/10.1016/bs.plm.2016.03.005.
Reinartz, W. J., Haenlein, M., & Henseler, J. (2009). An empirical comparison of the efficacy of covariance-based and variance-based SEM. *International Journal of Market Research, 26*(4), 332–344. http://dx.doi.org/10.1016/j.ijresmar.2009.08.001.
Rescher, N. (2009). *Aporetics: Rational deliberation in the face of inconsistency.* Pittsburgh, PA: University of Pittsburgh Press.
Rinehart, R. W., Duncan, R. G., Chinn, C. A., Atkins, T. A., & DiBenedetti, J. (2016). Critical design decisions for successful model-based inquiry in science classrooms. *International Journal of Designs for Learning, 7*(2), 17–40. http://dx.doi.org/10.14434/ijdl.v7i2.20137.
Sadler, T. D., Klosterman, M. L., & Topcu, M. S. (2011). Learning science content and socio-scientific reasoning through classroom explorations of global climate change. In T. D. Sadler (Ed.). *Socio-scientific issues in the classroom: Teaching, learning and research* (pp. 45–77). New York: Springer. http://dx.doi.org/10.1007/978-94-007-1159-4_4.
Salmon, W. C. (1994). Causality without counterfactuals. *Philosophy of Science,* 297–312. http://dx.doi.org/10.1086/289801.
Sandoval, W. A., & Reiser, B. J. (2004). Explanation-driven inquiry: Integrating conceptual and epistemic scaffolds for scientific inquiry. *Science Education, 88*(3), 345–372. http://dx.doi.org/10.1002/sce.10130.
Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of Educational Research, 75*(3), 417–453. http://dx.doi.org/10.3102/00346543075003417.
Smith, M. J., Southard, J. B., & Mably, C. (2002). *Investigating Earth systems: Climate and weather: Teacher's edition.* Armonk, NY: It's About Time Inc.
Stanovich, K. E. (1990). Concepts in developmental theories of reading skill: Cognitive resources, automaticity, and modularity. *Developmental Review, 10*(1), 72–100. http://dx.doi.org/10.1016/0273-2297(90)90005-O.
Stanovich, K. E. (2010). *Decision making and rationality in the modern world.* New York, NY: Oxford University Press.
Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Boston, MA: Pearson Education.
Tenenhaus, M., Esposito Vinzi, V., Chatelin, Y., & Lauro, C. (2005). PLS path modeling. *Computational Statistics and Data Analysis, 48*, 159–205. http://dx.doi.org/10.1016/j.csda.2004.03.005.
Tenenhaus, M., Amato, S., & Esposito Vinzi, V. (2004). A global goodness-of-fit index for PLS structural equation modelling. In Proceedings of the XLII SIS scientific meeting, Vol. contributed papers (pp. 739–742). Padova, Italy: CLEUP.
Van Merriënboer, J. J., Kirschner, P. A., & Kester, L. (2003). Taking the load off a learner's mind: Instructional design for complex learning. *Educational Psychologist, 38*(1), 5–13. http://dx.doi.org/10.1207/S15326985EP3801_2.
Wetzels, M., Odekerken-Schroder, G., & van Oppen, C. (2009). Using PLS path modeling for assessing hierarchical construct models: Guidelines and empirical illustration. *MIS Quarterly, 33*(1), 177–196. <http://www.jstor.org/stable/20650284>.
Willingham, D. T. (2008). Critical thinking: Why is it so hard to teach? *Arts Education Policy Review, 109*(4), 21–32. http://dx.doi.org/10.3200/AEPR.109.4.21-32.
Woodruff, D., & Wu, Y. F. (2012). Statistical considerations in choosing a test reliability coefficient. ACT Research Report Series, 2012 (10). ACT, Inc.